**Infrastructure for Collaboration in Support of High Performance Computing**

Michael E. Papka and Thomas D. Uram
Argonne National Laboratory

While computational resources and scientific datasets continue to grow relentlessly, the gap between scientists and their data and compute resources has grown accordingly. For scientists to efficiently use today's compute and data resources, a significant investment of time and effort is required; the coming increase in data and compute resource complexity as we transition to exascale will compound this problem many-fold. At the same time compute resources are becoming larger and more centralized, collaborations are becoming more frequently geographically distributed. To meet their research goals, scientists will require better facilities for managing their computational science campaigns, from planning, to code collaboration, to simulation, analysis, and discovery.

New analysis and visualization access models are needed that will improve utilization of these tools by scientists who are more often remote from their data, without requiring them to understand the details of the underlying compute and storage resources. New models are needed for querying and summarizing exascale datasets to improve the scientists' understanding of their data, while concealing the details of the underlying storage resources. New methods are needed to track the activities of scientists during campaigns into their data, to help them to record their progress so they can later reproduce those actions to produce the same results, and to apply them to related datasets. New methods are needed by which scientists can share their data, workflows, and results with their collaborators and their larger communities, where they can be reused to produce new insights in the combined space.

A framework is needed that will bridge the gap between scientists and their computational science, simplifying the process generally and automating it where possible. Using this framework, scientists could easily describe their data using domain-specific data models, such as numerical variables of simulations and timestamps as well as file schemas. They could design analysis pipelines using script languages and workflow toolkits that understand their domain-specific data models. These scripts would embody the computational science workflow, and be used by the framework for generating user interfaces and for managing inputs, execution, outputs, and derived data products. In this way, the framework would provide a uniform interface for managing the scientific workflow, supporting distributed collaborations with a set of common tools and a pedigree for their simulations.

The overarching need is for a collection of components from which scientists can assemble environments for conducting their science in the currency of their domain, without requiring them to understand the details of the underlying data and compute resources, even as these relentlessly expand in size and complexity. Scientific productivity will be enhanced by enabling several key functions: browsing of datasets in domain-specific semantics rather than filesystem semantics; simple access to a suite of visualization techniques, flavored by recommendations from colleagues and from the system; provenance capture and browsing to track one's progress, to reuse work, and to reproduce results; and synchronous and asynchronous collaboration with colleagues around data, workflows, visualizations, and their products.

While each science has its idiosyncrasies in analysis and visualization applications and methods, some functions are common between sciences; a general environment can provide these common facilities, and enable domain-specific customizations. Annotation, query, monitoring compute resources, monitoring running jobs, viewing job history, and managing data can be captured and exposed as the core of a system supporting tomorrow's science.

An example that shares commonality with many other DOE facilities (e.g. NERSC, APS, SNS) is the DOE Leadership Computing Facilities located at Argonne National Laboratory and Oak Ridge National Laboratory. These facilities provide computational resources to a global community of users each year. In 2012, a total of 2.7 billion hours of CPU time will be provided, with 1.7 billion of those hours going to the INCITE (Innovative and Novel Computational Impact on Theory and Experiment) program. In 2012, the INCITE program will support 60 projects with membership distributed across more than 60 different locations. These projects will generate hundreds to thousands of terabytes of data that needs to be analyzed by scientists that are not collocated with the data; in fact only 30% of the projects have a member at either facility. This means 70% have no members collocated with the projects data and will need to interact with the data remotely. Furthermore, each project on average has more than two institutions involved adding the need for collaboration support between the partners and the data.